Universitat de Lleida

TREBALL FINAL DE GRAU

ESCOLA POLITÈCNICA SUPERIOR
UNIVERSITAT DE LLEIDA
INSPIRING THE FUTURE

Estudiant: Marc Sances Medina

Titulació: Doble Grau en Enginyeria Informàtica i ADE

Títol de Treball Final de Grau: **Design of Automated Trading Systems for Foreign Exchange Markets**

Director/a: **Fernando Cores Prado / Anna Vendrell Vilanova**

Presentació

Mes: Setembre

Any: 2019

Abstract

The purpose of this research is to offer an initial approach to the design of automated trading systems, by analyzing the current state-of-the-art, identifying the available financial markets and their nature, and implementing a proof-of-concept which operates with a variety of statistical techniques.

Author

About the author	
Name	Marc Sances Medina
Faculties	Escola Politecnica Superior (Polytechnic School) Facultat de Dret i Economia (Law and Economy Faculty)
University	University of Lleida
Degree	Double Degree: Computer Engineering and Business Administration and Management
E-mail	msm21@alumnes.udl.cat
About the directors	
Name	Fernando Cores Prado
Faculty	Escola Politecnica Superior (Polytechnic School)
Field of knowledge	Computer Architecture and Technology
E-mail	fcores@diei.udl.cat
Name	Anna Vendrell Vilanova
Faculty	Facultat de Dret i Economia (Law and Economy Faculty)
Field of knowledge	Financial economy and Accounting
E-mail	ana.vendrell@aegern.udl.cat

Contents

Chapter 1

Introduction

In this chapter, the motivation of the project is explained. The scope will be defined and the objectives and timing will be shown. Finally, the structure of this report will be detailed.

1.1 Motivation

The information era brings us new challenges in every single field of knowledge. And the financial markets are no exception. Indeed, these markets are extremely dependant on information, and thanks to the information technology, they work faster and thousands of operations can happen in one second. The stock markets are extremely competitive, and being one step ahead the competitors can mean large benefits. So the development of systems to assist or even replace broker's function is becoming more necessary in the current era.

1.2 Scope

The project will be focused in studying the currently available technologies for stock prediction, as well as the financial markets available for trading. After such study is performed, a proof-of-concept will be designed which performs decisions in the chosen market.

1.3 Objectives

- Design of a system able to perform intra-day operations in the financial markets and detect changes in the trends, and react in consequence.
- Finding of the right balance between profitability and risk in the system.
- Design of a scalable system that can operate in large clusters effectively.
- Comparison between the implemented system and the current state-of-the-art.

1.4 Timing

This project will be timed in three phases: an initial research, the implementation phase, and a testing and validation phase. The goals on each phase are detailed as follows.

1.4.1 Initial Research

In the first phase, a research on the current state-of-the-art will be performed, alongside with a feasibility study of the proposed alternative to be implemented. The tasks in this phase will be:

- Gathering of stock data, with as much precision as possible.
- Study of the stock markets, to find feasible choices for setting the scope of the project.
  - Election of the markets against which the system will be tested.
  - Election of the frequency of investment of the system (intra-day or not, hour, minute, second or tick precision).
- Study of the state-of-the-art, research on the current ATS available in the market and their features, alongside with their reliability and profitability, and the technical challenges currently existing in this field of research.
- Study of the available Big Data platforms and systems, and election of an appropiate system for the purpose of the project.

1.4.2 Implementation

In the implementation phase, the system will be implemented.

- Election of the appropiate algorithm or algorithms to detect trends in the stock market and operate.
- Design and implementation of the system with the chosen technology.
- Initial tests with sample data.

1.4.3 Testing and validation

In the last phase, both technical and economical tests of the system will be conducted, in order to further test the feasibility of the system.

- The technical study will test and validate the correct operation of the system in all conditions.

- The economical study will study the economic variables of the outcome with the system. The system will be tested against real data, testing the economic variables that show profitability and risk of the overall system, and comparing it to the current available systems.
- Finally, the outcome and conclusions of the research will be presented.

1.5 Structure

This project is structured in the following sections:

- **State-of-the-art of ATS** describes the current state-of-the-art of the automated trading, giving an introduction to the high-frequency trading paradigm and describing the techniques currently available.
- **Financial markets: where to invest?** explains the nature of the financial markets, describing the main financial instruments available and exposing their benefits and risks.
- **Methodology** defines the methodology employed to design a proof-of-concept and the strategy that will be executed.
- **Development** details the technical process of development, and the different iterations over the design of the software.
- **Testing** shows the outcome of the execution of the proof-of-concept and the results of its performance in the market.
- **Conclusions** extract final conclusions and call for future work on this subject.

Chapter 2

State-of-the-art of ATS

This chapter will cover the current state-of-the-art, detailing the available ATS algorithms and techniques and the current Big Data platforms and their features.

2.1 High-frequency Trading paradigm

The introduction of new legislation in the US, with the Regulation National Market System law (Reg NMS, 2005) and in Europe, with the Markets in Financial Instruments Directive (MiFID, 2007), defined the beginning of a new paradigm, where High-frequency trading strategies (HFT) began being more profitable and useful, and rapidly started gaining interest among investors. [Easley et al., 2015]

The new legal paradigm allowed for greater market fragmentation, therefore increasing competitiveness and rewarding innovation and technological improvements.

2.1.1 What is HFT?

In essence, High-frequency Trading is a trading paradigm which focuses on executing orders by analysing the available information in the market and making a decision over buying or selling a given amount of a commodity at a given price, quickly reacting to market changes as they introduce. At a first glance, it might seem that HFT always performs massive amounts of operations every second. However, time is not a variable at all for HFT strategies, as HFT orders are triggered by the changes in the market. Therefore, HFT can adapt to the pace of the rest of the agents in the market, and speed up or slow down based on the volume of the operations. [Easley et al., 2015]

This kind of trading strategies have already shown weaknesses. In May 6th, 2010, US equity markets plummeted to rapidly recover by the end of the day, in an unprecedented "flash crash", affected by a sell order from a single trader that ignited reactions from high-frequency traders. This shows that HFT algorithms can operate as fast as their feedback allows them, and for such reason, they must be carefully planned and designed.[Akansu, 2017]

2.2 Current ATS algorithms and techniques

Currently, there are several ATS algorithms, each one with different features. Shobana and Navale show comprehensive reviews on the available systems. Primarly, Artificial Neural Networks are the most accurate systems in predicting stock markets, but there exist several approaches in performing this analysis and there are other algorithms which don't make use of ANN.

[Shobana and Umamakeswari, 2016] [Navale et al., 2016]

2.2.1 Artificial Neural Network (ANN) based systems and variants

This category of ATS algorithms covers systems which use artificial neural networks (ANNs) in order to analyze the data and make predictions.

Though they provide a with self-reinforcing model that improves over time, and they can detect hidden patterns than other algorithms do not reach, they require high amounts of computing power and a long and complex training with huge amounts of data. This makes their economic cost significantly higher than the other alternatives. They are profitable only after being provided with a large amount of traning data and long time and processing power to train their models and improve their precision. [Kara et al., 2011]

2.2.2 Genetic algorithms

This category of ATS algorithms covers self-reinforced algorithms which successfully learn from historic data by evolution and mutation. Whilst the ANNs try to model human neurons and simulate the behavior of the human brain, genetic algorithms imitate the behavior of the genetics and evolution. Each iteration of the algorithm, a population of solutions to the problem is breed, and evaluated against a certain fitness function that selects the best alternatives among the population. Such alternatives survive to the next iteration by either crossing them over (combining two existent solutions), and mutation (occasional random changes in the parameters to guarantee diversity). They are less complex than ANNs, but still require large amounts of data and resources.[Armano et al., 2005]

2.2.3 Support Vector Machine (SVM) based systems

Support Vector Machine systems provide with an algorithm that can build classifiers. Given a certain input, the algorithm can tell the class the given input belongs to. In a financial context, this can be a "high vs low" classification, where the inputs are the values of the stock and a decision is given on which direction will the asset turn, and then make the appropiate order. However, SVMs cannot easily predict specific values of the asset, only their direction, making their usage rather limited.[Kara et al., 2011]

2.2.4 Econometric and statistical approaches

This category of ATS use more traditional prediction models, like regression models, in order to predict future values of the assets. However, since the output functions are much simpler and have less coefficients than the previously cited algorithms, they are less precise and more prone to not responding to uncommon patterns. Despite that, their computational and economic cost is significantly lower, and for such reason this systems are also popular albeit less profitable.[Kirchgässner et al., 2013]

This is the chosen methodology to design our proof of concept.

2.2.5 Environmental Analysis

This category of ATS focuses on the analysis of the environment related to the markets, by analyzing press and sentiments to predict market trends. It is common that big announcements in companies are responded with a movement in the stocks, so analyzing this information as it comes by and responding accordingly can be a feasible strategy to predict patterns that cannot rely solely on the historic values of the stock.

However, the subjectivity of the data and the lower availabilty of inputs in this type of analysis implies that this algorithms may be used only to complement the behavior of other algorithms that do read the stocks. [Attigeri et al., 2016]

2.2.6 Data Mining

This category of ATS uses data mining techniques to find patterns in large amounts of historic data, and use the patterns to make predictions. By analyzing unusual data records (for example, an uncommon amount of activity in a stock), relationship among variables (like the relationship between the market values of several stocks), grouping data into similar clusters, classifying data (high-low), building regression models and summarizing the information, it is possible to find a strategy to predict future values based on identifying the patterns and their expected output.[Navale et al., 2016]

Chapter 3

Financial markets: where to invest?

Understanding of the financial markets and the available choices for investment is essential to design a strategy to take maximum profit of the market characteristics. This chapter will cover the equity and currency markets, and which ones are better when performing investments. Afterwards, based on all the research done, a decision will be taken on which ones are chosen for the research.

3.1 Financial market key concepts

Before getting started with a description of the financial markets available, it is important to emphasize the key concepts used in this chapter.

When operating in an exchange, instead of a fixed transaction price, two prices are registered, which are bid and ask. On the one hand, the bid price tells the maximum price a buyer is willing to pay for a commodity, whilst the ask price is the minimum price a seller will accept. In this situation, the market is stalled: none of either parts will confirm the operation. The urge for either buying or selling the asset (or both) will lead buyers and sellers to agree on a transaction price and closing the operation, then the next bid and ask prices are shown. [Investopedia, 2018]

The spread among both prices will explain the asset liquidity and market activity. If the spread is very low, the market will likely have more activity, because buyers and sellers will be closer to a position where an order can be executed. However, if there is a large difference between both prices, it will be less likely that an order is closed, because neither the buyer will accept the high ask prices, nor the seller will sell at the low bid prices. Therefore, less operations will be executed until this spread reduces.

When deciding in which market to invest, it is mandatory to study certain aspects of the market that immediately affect profitability.

Volatility of a market defines how often and how much the price of the asset will change. A highly-volatile asset, such as cryptocurrencies, can lose all of its value in a minuscule time span, which means volatile assets have a higher risk profile. While an automatic trading system should be able to detect volatility to a certain measure, it is important that a safer asset is chosen in order to prevent losses.[Investopedia, 2019]

Markets are also subject to speculation. Speculation involves buying and selling assets in large amounts for very short amounts of time, where the price can be predicted with a high degree of certainty. While this gives small profits, over time it can be a successful strategy. However, speculation causes massive amounts of the asset to be sold or bought in this short time spans, which may even affect the asset price. While speculation is present in most markets, it is more likely to happen in volatile markets.

3.2 Equity markets

Equities have been subject of trading since the 17th Century[Beattie, 2018], and they're traded both by small, medium and large investors, in endless combinations of volume, profitability and risk.

An equity is "what a shareholder owns in a corporation, entitling him/her to part of that entity's profits (in the form of dividends) and a measure of control (through shareholder voting rights)."[Financial Times,]

However, equities are not just about dividends: company shares have a market value. Initially shares are just a participation of the company's capital, but once they start to be negotiated in secondary markets, they build up a market value that will be altered by news, payout expectations and company results. This creates an interesting opportunity for traders that just require profitability instead of participation in a company.

3.2.1 Benefits of equity markets

Company shares are one of the simplest financial instruments available. Profitability can be obtained by taking advantage of price differences over time, or by the payment of dividends. Company benefits may not be fully paid, but reinvested into reserves. This decision, if correctly executed, can also improve the profitability of the shares, as the investments made may help get larger benefits in the future, increasing market expectations and therefore multiplying the stock value of the company.

3.2.2 Risks of equity markets

Companies can go bankrupt, and shareholders are among the last ones to be paid. As data is controlled by stock exchanges, it's hard to find historical data, and the data available is not cheap.

3.2.3 Evaluation of risk for equity markets

When we have to work in an equity market, it is important to study the risk of the company involved, since it can help predict the expected value of its stocks.

Gorbunova ([Gorbunova, 2016]) mentions the traditional methods for stock analysis. Among its choices, there is the top-down and bottom-up multilevel

forecasting. This method focuses on studying the economic situation of both the company subject of trading all the way up (or down) to the situation of the country where the company operates. With the top-down method, we look to infer projections of the expected company growth based on the sector and national economics forecasts, whilst with the bottom-up method, we look to translate the situation of the company into forecasts of the country economy.

Gorbunova makes a remark that the bottom-up strategy is less correct, since the growth or decline of a single company is not relevant enough to influence the situation of a country. However, combined with the top-down strategy, it can provide a much better picture of the situation of a company.

The three stages of this multi-level analysis are a market conjuncture analysis, a macroeconomic analysis and a microeconomic analysis.

The market conjuncture analysis is a general analysis based on the combination of both microeconomic and macroeconomic factors, focusing on the dependence between them and the effects they have on the supply and demand.

On the other hand, macroeconomic analysis focuses on nation and region-wide statistics that can explain the situation of a country's economy. Among this, the most standard macroeconomic indications are the GDP (gross domestic product), inflation, unemployment rate, debt, turnover and exchange rates. However, macroeconomic analysis depends on a large number of variables, and even when properly forecasted, its impact on the company expectations is more indirect.

Half the way between macroeconomics and microeconomics, we can study the risk of the economic sectors where the company operates. Knowing the sector productivity, sales, prices and expectations of growth and comparing them to the company figures will show whether the company is operating properly in its industry and whether the industry where its operating has positive or negative growth expectations.

Finally, microeconomic analysis evaluates the company's status in the stock market and financial situation. This microeconomic analysis can be performed using indicators of the company's market activity. The most usual ratios being:

- **Earnings per share (EPS):** It is an indicator of the attractiveness of a stock, as it determines the amount of profit a company can generate for its shares. It is calculated by dividing the net income (excluding

dividends) by the available company shares. [Investopedia, a]

- **Price Earnings Ratio (PER):** It is a more specific indicator of the EPS ratio, since it divides the market value per share by the earnings per share itself. The result of the operation is a ratio that can be compared between companies as an indicator of the performance of the company stocks.

- **Earnings Ratio Price (PER inverse):** It shows profitability of a stock according to its market value. While PER is more used for comparisons, the inverse of PER is valid for evaluating company's ability of generating benefits.

- **Dividend Payout:** It is a calculation of the dividends by the net income. This shows the divident payout policy of the company and the returns of the company to its shareholders as dividends.

- **Market-to-book Ratio:** It evaluates the price of a company in the market versus its accountable assets. As a company has higher growth expectations or a more positive forecast, investors will agree to pay more for the company than its current asset value, since they expect this assets to grow as the company gets more benefits. However, it can also become a red flag for a company that is overvalued in the market and such, its stocks will be less appealing since they will be too expensive to purchase.

- **Price to sales ratio:** It shows the amount of money investors will pay for each unit of sales. This ratio can be used for comparisons between companies, showing the stocks that generate less sales for each dollar invested.

The main issue that traditional multi-level forecasting presents to the purpose of this research is that it requires interpretation and subjective analysis of the company situation, which is a hard to automate process and can lead to bias. Therefore, an alternative approach for studying the risk of equity markets is required. Even though microeconomic indicators are feasible to evaluate automatically, through an analysis of the company accounts, they still require to be interpreted in a more global scope rather than being individually evaluated.

Therefore, an alternative approach for studying the risk of an individual company, that is objective and automatable, is required if we require operating in stock markets.

Price analysis, i.e. comparison of the historic prices of a stock and the relationship with other company's stocks, can be automated and predicted in the short term. However, it is subject to the market volatility, and equity markets can be seriously affected by the previously mentioned factors, which do not depend on the price historic. Currency markets, however, are less prone to this volatility, since they only depend in macroeconomic factors, and not particular industry or company situations. For such reason, the following section will study the currency markets in depth to evaluate the feasibility of predicting them.

3.3 Currency markets

Foreign currency markets, also known as Forex (**For**eign **Ex**change), have also played an important role on the global capital market for decades. In Forex markets, foreign currency is the main asset subject of trading.

The financial instruments for foreign exchange are:

- Spot trading: it's the fastest forex trading instrument. In most spot currency trade contracts, the exchange is made effective two business days after the execution of the trade, with the exception of the US dollar vs Canadian dollar, which is done the next business day. It's important to notice that the settlement day must be a valid business day in both currencies, i.e. a national holiday in one of both countries or regions might delay the effect of the operation.[Investopedia, c]

- Forward exchange: the quantity and price is decided on the signing of the contract, but the settlement is produced in a later date. Forward contracts differ from futures in that they are freely negotiated between the interested parties and not traded in a standardised format and a controlled, central market. Therefore, the default risk is higher.[Investopedia, b]

- Financial derivatives: treated in more detail in the next section.

3.3.1 Benefits of currency markets

The information of currency markets is more open and available, with on-line services offering free intraday information, with precision ranging from the hour to the tick.

The analysis of currency markets can heavily rely on historic prices and comparison of the performance of other currencies, unlike equity markets, since currency markets are less volatile and its price depends on less factors than the stock of a company.

3.3.2 Risks of currency markets

On the other hand, foreign currency is more complex than company equities. It is more common to trade forwards and futures, and even spot contracts are not immediate, and therefore margin is lower. The execution of an order cannot be reversed until at least two business days, so an abrupt change in the market might be harder to recall. However, all participants of the market play with the same rules, so it doesn't matter that much when the actual operation is settled.

3.4 Financial derivatives

Derivatives are contracts that are based on the performance of a certain financial instrument, asset, index or interest rate. The most common financial derivatives are forward contracts, futures, options, warrants and swaps. Their main basis relies on exchanging assets or their cash flows in a future moment, at a price agreed in the moment of the purchase.

3.4.1 Benefits of financial derivatives

Financial derivatives give an outstanding profitability if the investor has an advantage over the market and can predict the market trends before they actually happen. They provide more financial leverage to the investor, which can profit from the future outcomes of an asset without having to hold it for the time it takes to generate the profit.

3.4.2 Risks of financial derivatives

However, financial derivatives cannot be recalled early (except american options). The owner of a derivative must wait for the settlement of the operation to take action on it. Moreover, futures and forwards can be a huge loss to their owner, since it is enforced to perform the transaction, unlike options (where it could not perform the trade if it was lossy to him).

Financial derivatives are complex and traded in less informed markets and exchanges. The fact that most of them are over-the-counter makes them even less appealing to the investor that looks for an automated solution. They also prove harder to analyze, since they are a commodity on their own with its market price, and rely on another underlying asset, so its analysis requires knowing when to buy the contract and when it should expire.

Derivatives are set in a tiny gap between an investment and gambling, since they rely strongly on the ability of the trader to predict the future value of the assets.

3.5 Cryptocurrency markets

Cryptocurrencies have become a popular instrument for trading in the last years. Although they have not developed completely as a primary currency, they are a heavy subject of speculation, with high volatility and strong fluctuation. Cryptocurrencies work by securing the transactions using a cryptographic technology known as blockchain, making the transactions extremely hard to tamper with the currently available computers.

The largest and most known cryptocurrency, as well as the first effective implementation, is Bitcoin, implemented in 2009 by Satoshi Nakamoto (possible alias, identity unknown). Bitcoin has been subject of study for its lack of regulation and central market.[BARTOS, 2015] [Gervais et al., 2014] [Yermack, 2015]

3.5.1 Benefits of cryptocurrencies

Cryptocurrencies are an open system where transactions are known by all the peers and verified by a handful of them, so there's no financial secret on the operations. However, external exchange operations (i.e. "standard" currency vs Bitcoin) are done by private exchanges that may or may not pub-

licly acknowledge the operations and exchange rates. Either way, exchange information is more or less as available as in Forex markets.

3.5.2 Risks of cryptocurrencies

Cryptocurrencies have been subject of controversy because of the illegal uses, lack of regulation and high energy consumption leading to a great environmental footprint. There is no central exchange, although some of the largest exchanges might run in oligopoly. Moreover, the developers of the Bitcoin software have a large control over the currency and might commit fraud by tampering the protocol internals for profit. So, despite crpytocurrencies are controlled by IT developers and owners of mining pools, rather than banks and governments, there is indeed certain control over the currency.[Gervais et al., 2014]

Volatility is also an important drawback on cryptocurrencies. While in traditional foreign currencies the governments can impact the price of the money by using monetary policy instruments such as interest rates and quantitative easing, cryptocurrencies are not subject to any authority, and their price depends exclusively on the market.

This means no short-term measures can be taken in order to stabilise the price of the currency, and a small fluctuation in the market may trigger a massive reaction, dramatically rocketing or plummeting the value of the asset at a very fast pace.

There is as well a high amount of speculation, due to the high short-term fluctuations in the asset value, which allow to make small profits over time by buying the asset when a rise in prices is predicted, waiting for such rise and then selling, and repeating the process across several assets.

3.6 Project Requirements

This project demands information availability over other factors. For this reason, the main priority has been to find precise, tick data of several markets. Foreign exchange (FX) tick data has been the most available of all the markets. Some clearance houses offer historic tick data, which is less available for stock trades.

Although cryptocurrency data is widely available as well, it will not be used as the target trading asset because of its high volatility, which makes it

more unpredictable due to the less strict regulation.

Equity market information is hard to find. Unlike foreign currency, which depends on several clearance houses, equity stock information is controlled by the major stock trades, which charge for the most precise information (although they do provide basic level information with some delay). In Spain, the main stock trade only offers daily information. Some non-primary financial information sources do provide this information with minute precision, but only for a short time span, with no option to download minute information of several years. Some historic tick data was found, but only for equity CFDs, which are not the same as liquid shares, and are negotiated in over-the-counter exchanges.

Therefore, the final choice for this project is to study foreign exchange (FX) spot trading historic data, although it has been considered aggregating information from other markets to study the effect over the target currency market value.

Chapter 4

Methodology

This chapter will cover the methodology for development of the ATS system, economical and technical testings and the data sources used for the study.

4.1 Aligning data for the linear regression

The market information comes in discrete time spans, where there can be no operations in several hours and large amounts of movements in a few seconds. Due to this heterogeneous temporal distribution of the data, performing a regression of this information requires it to be transformed (aligned) so that every value of the input for the regression represents the market value of the asset in a fixed time frame (for example, every 15 seconds).

To perform this data alignment, the copy-down method is used. This method works as follows:

- A time frame for diving the data in chunks is determined. This can be based on the availability of information, the shorter the time frame, the more precise the model will be, however, if the time frame is faster than the time required for events to occur, then it will add unnecessary computational cost to the problem.
- For each time frame, if there are one or more events (prices) available, the average of all prices is calculated and assigned to the time frame.
- If there are no events for a certain time frame (for example, at night with the markets closed), the value assigned to the time frame will be the average between the value of the previous time frame, and the value of the next time frame. This way, the information gaps are filled with averages that do not skew the data model.

This method can be easily implemented by scripting, however, it has proven hard to implement in a parallel computer cluster due to the nature of the map-reduce paradigm. Since the data is distributed among a large set of nodes for processing, and the nodes can't access other node's data easily, filling the gaps becomes challenging.

Therefore, the approach taken to align the data in the implementation considers that the amount of time required to process the historic information and align is an acceptable delay even when it's not processed in parallel, since it is an operation that will only be performed once. For such reason, the training will be performed before operating, and the resulting model will be reused.

4.2 Mechanisms for regression

In order to perform predictions, there are different statistical approaches that can be taken into account.

4.2.1 Single variable regression models

Autoregressive model

Autoregressive models are among the most simple models for prediction in time series. They define that the future values of a time series depend linearly on its previous values. Namely, the formula for an autoregressive model of a single lookahead term is defined by the equation 4.1.

$$x_t = \delta + \alpha x_{t-1} + \epsilon_t \tag{4.1}$$

Where x_t would be the forecasted price at moment t, δ is a constant value, α is a coefficient that will multiply the price at moment $t-1$ and ϵ_t is the error term.

Autoregressive models can be easily implemented using ready available libraries such as statsmodels and predictions can be made from such models, so it is the chosen start point for the proof of concept.

Among its drawbacks, the time windows should be narrow enough in order to predict trends correctly, since a large training set will not be meaningful to the analysis due to the presence of long-term trends that are not influenced by the values in short time spans.

It can't either find correlation among symbols, since it works with a single time series, making the prediction rather simple and less precise than other more complex models. [Kirchgässner et al., 2013]

Moving-average model

Moving-average models take into account the error of the forecasts rather than the actual value found. This affects indirectly the future forecasts. This model shows the advantage of reducing the impact of short term trends in the market, but it's less responsive in short term. A moving-average model of a single lookahead term would be defined as shown in equation 4.2.

$$x_t = \mu + \epsilon_t + \theta \epsilon_{t-1} \tag{4.2}$$

[Kirchgässner et al., 2013]

Being x_t the forecasted price at moment t, μ the mean of the time series, ϵ_t the error term for the current moment, and θ the coefficient which multiplies the error term for the previous moment.

Unlike autoregressive models, moving-average models have a limited impact in future forecasts, since predictions depend on the average and a finite amount of error terms, and not in the previous forecasted values (which could propagate error over an infinite amount of time). [Kirchgässner et al., 2013]

Autoregressive moving-average model (ARMA)

The combination of the autoregressive and moving-average models yields the autoregressive moving-average model (ARMA), which contains both models.

An ARMA(1,1) model is defined by the formula in equation 4.3.

$$x_t = \delta + \epsilon_t + \alpha x_{t-1} + \mu_t - \beta \epsilon_{t-1} \tag{4.3}$$

Where x_t is the forecasted price at time t, δ is a constant value, αx_{t-1} is a coefficient multiplied by the price at time $t-1$, and ϵ_{t-1} is the error term for time $t-1$.

Autoregressive integrated moving-average model (ARIMA)

Finally, further improvements over the ARMA model yield the ARIMA model, which reduces the impact of the level of the data by computing the differences of the data in order to make it stationary, predicting the trend of the series rather than forecasting its value.

4.3 Mechanisms for operation

Once predictions are found, the system will have to operate in a live environment, taking decisions based on the forecasted prices of the symbols. Such decisions are called orders and will be defined by the trader.

4.3.1 Market orders

Market orders are immediate orders to buy or sell an asset. They are executed at the current price of the commodity, without waiting to any condition. [U.S. Securities and Exchange Commission, 2019]

4.3.2 Limit order

Limit orders require to buy or sell an asset at a certain price or better, keeping the order in hold until the desired price is reached.
[U.S. Securities and Exchange Commission, 2019]

4.3.3 Stop-loss order

Stop-loss orders require to buy or sell an asset when a certain stop price is reached. It differs from the limit order in that the order will be executed unconditionally and immediately (like a market order) once the specified price is reached. [U.S. Securities and Exchange Commission, 2019]

4.3.4 Forecasting market orders

The ability of forecasting the market trend and predicting future values on assets can allow the trader to define more specific orders, where the expected value of an asset is take into account in order to automatically place market orders. A rising trend on an asset will allow to place a limit sell order in order to gain maximum profit, and detecting the decay of such trend timely can allow to execute the order while profit can still be reached.

4.4 Data sources used

Obtaining proper data sources is one of the greatest issues. It's hard to find precise tick data, even for historical data, as such information is used often for ATS testing and development, and there is commercial profit behind this data.

The most accessible data which accomplished the research requirements has been foreign exchange ("Forex") data, which is provided free of charge at Pepperstone website[Pepperstone,], an Australian Forex clearance house. The data is available for several common currencies (Australian dollars - AUD, Canadian dollars - CAD, Swiss francs - CHF, Japan Yens - JPY, Euros - EUR, British pounds - GBP, New Zealand dollars - NZD and US dollars - USD), with tick precision, from May 2009 to November 2016. This data source is already large enough for the purpose of this project, so it has been chosen as the main data source for testing.

The format of the data is Comma-Separated Values (CSV), which means a text file is provided where each line represents a register of data, and the values of each column are separated using a comma (,) symbol. This format is simple enough to allow parsing large amounts of data with high performance, since it's the format with less data overhead, whilst still readable enough.

The data is split month by month, for each pair of currencies. Each of the files is a CSV file, compressed in ZIP format for space reasons. The CSV file contains the tick information, line by line, in format CURRENCY, TIMESTAMP, BID, ASK.

Chapter 5

Development

The following chapter details the development progress. The main issues found will be detailed, the project structure and the initial tests will be covered as well.

5.1 Details of the project

For the development of the proof-of-concept ATS that is developed in this project, the following technologies are used.

- Python 3.7 with numpy 1.16, scipy 1.3 and pandas 0.24, deployed as a Docker image.
- The actual tests will be run without Docker image, in a cluster property of the faculty.
- The project is deployed using Git and GitLab.

5.2 First attempt in Spark

The original approach to the project was using Apache Spark. Apache Spark is a distributed analytics engine which allows processing of large amounts of data, so it showed quite capable for this situation.

However, several struggles have been found in this approach. The main issue comes when aligning the data. As it has been explained in section 4.1, the data is provided in a times series format with irregular tick distribution, i.e. the data comes at unknown timespans.

In order to work with such time series, it is required to perform an alignment of the data, which will provide evenly distributed price information

Performing this alignment can't be easily distributed among a cluster of machines, because the alignment process may require information that is being processed in another node.

Therefore, after several attempts to approach this issue, Spark was finally discarded for its complexity, providing a simple proof-of-concept using the SciPy stack.

5.3 Initial iteration

A first program to test the development environment is run. The test project takes a CSV file named "sample.csv" which contains foreign exchange information from our datasets, in the format CURRENCY, TIMESTAMP, BID,

ASK. The program computes the average open and close value of each currency type, by loading the dataset into a Pandas data frame and then reducing by currency type (EURUSD, EURGBP, etc). The reduction function, given two tick lines of the same currency type, casts the BID and ASK values to float, computes the average between the two ticks, sets the TIMESTAMP to AVERAGE and keeps the CURRENCY type, returning a tick line without timestamp and with the BID and ASK average of the two tick lines.

This program is simple enough to test numpy while it works already parsing the CSV files that will be served as input, so it can be run to test all the storage mechanisms (local, S3, NFS, etc).

5.4 Strategy selection

The proof of concept should be able to switch between several prediction algorithms, which have been described in the Methodology section (4.2), so that a comparison between each algorithm can be performed in order to extract conclusions.

In order to achieve this modularity, the Strategy software design pattern is suggested. In the Strategy pattern, the classes that need to perform a certain action will not know about how such action will be implemented, and will instead call a generic "strategy" interface and ask it to perform the action. This allows to change at any time the concrete implementation of the strategy that will be used.

While Python has a less strict object-oriented design than other object-oriented programming languages such as Java or Scala, the base concept of the design pattern can still be applied, although not enforced.

The figure 5.1 shows how the Strategy pattern design is achieved, by designing an abstract IPredictor which acts as an abstract base class, and then implementing multiple specific predictors which inherit from such abstract class.

Thanks to the use of this pattern, the strategy is chosen in runtime based on the user request, and the same global implementation of the trading system can be reused across multiple prediction strategies.

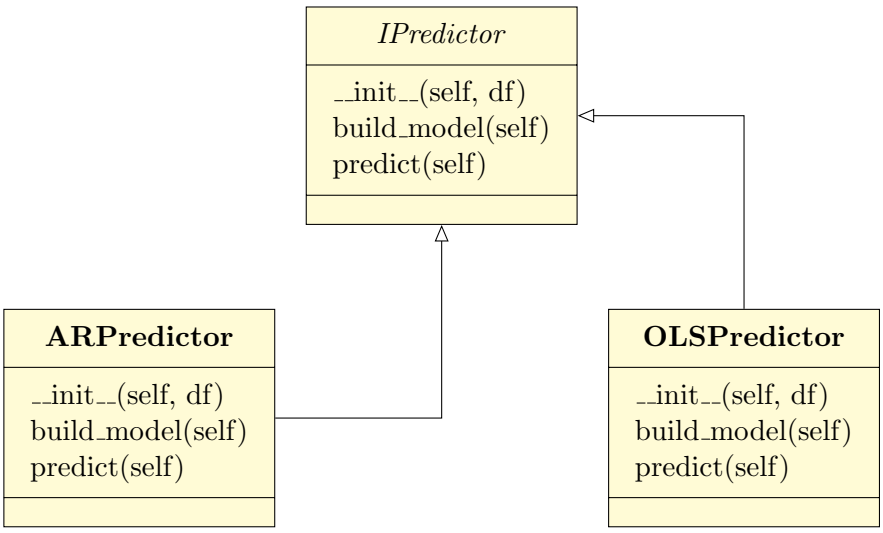

Figure 5.1: UML schema of the Strategy pattern used

5.5 Providing a live stream of information

Another technical challenge in this project is being able to process the data in real time. While executing orders is out of the scope of the system developed in this project, the system should be able to work with a live stream of information, making predictions and taking decisions in basis to such information received.

The strategy chosen for working with live information streams is just reading the information from standard input, which can be piped from an external software, although reading from a file is also allowed.

The program makes use of the system call *select* to wait for input to arrive through the stream. As the data arrives, it is temporarily collected in memory (filesystem is suggested if the datasets become too large to process), until a reasonable amount of data (namely "update window") is reached.

When the update window is reached, the model used for prediction is calculated again in a separate thread. Threading allows to keep operating in real time while more data is getting ready for processing. Once the new data is processed and a new model is available, the old model is replaced with the new one, which will be more precise.

Another small challenge is being able to tell the program to stop reading

information, since it does block waiting for more information. Although it was considered an important deal when working with finite sized datasets, and an approach using end of file (EOF) markers was made, since the real-time situation will likely work for an undetermined amount of time, the simplest way to stop the processing is just by sending the termination signal (SIGTERM) to the process.

5.6 Deployment and usage of the project

In order to deploy the project, two mechanisms of deployment are provided: Python virtual environments (virtualenv), and Docker images.

5.6.1 Version control with Git

Git is a distributed version control system, which is a software designed to keep track of changes to code, by keeping an historic of the changes made to the code, which are known in Git as *commits*, and allowing to isolate the changes in branches, allowing teams to concurrently work in new features.

Despite the development of this project is done by a single person, Git is being used as a tool for keeping the code safe in a remote location and for working in tentative changes that might need to be rolled back.

Work in progress is done in a separate branch, usually "wip" or "feature", where new code is implemented. Once the code is ready, the branch is merged into the main (master) branch. If the code eventually renders nonfunctional, restoring a backup is as simple as checking out the master branch, which contains the stable code, and starting over again with new code. [Chacon and Straub, 2014]

5.6.2 Virtualenv

Virtual environments are a non-portable system that allows to create isolated Python environments, where libraries are installed in a non-system wide manner, so that version conflicts can be avoided and the exact requirements of the application can be met.

Deployment using virtual environments requires creating the virtual environment, installing the project requirements (provided using a require-

ments.txt file) and then activating the environment using the provided activation script (bin/activate).

Upon activation, both the Python interpreter and the packages of the virtual environment will be available for any Python script running in the environment. [Virtualenv,]

5.6.3 Docker image

Whilst Virtualenv can isolate the packages, scripts and environment variables of the OS, it is not a formal virtualization system that can securely isolate the applications from each other.

Docker resolves this issue, providing a powerful OS-level virtualization, where the filesystem is completely isolated to a so-called "container", however, unlike conventional hypervisors, it shares the operating system kernel (which must be Linux, or accordingly emulated). This makes Docker containers extremely lightweight, allowing massive scaling and deployment in a microservice structure. [Docker, 2019]

Containers are also portable, meaning they can be quickly replicated and deployed to any system just by installing Docker and pulling the image.

Furthermore, Docker defines the build steps of the images in a plain text file, the Dockerfile, which reduces the complexity of the mantainance by defining simple steps of the build process.

A trading system to be deployed to several machines will likely benefit from this portability of Docker images, with a rather acceptable impact on performance. For this reason, a Dockerfile has been included in the project, which allows the project to be compiled and deployed with less complexity than Virtualenv.

In order to build the project, we just need to run the following command:

```
docker build -t tfg-project:latest .
```

Which will build the Docker image, running each of the steps defined in the Dockerfile, and give it a recognisable tag called *tfg-project*.

Then, the image can be saved to a tarball, i.e. a single file containing all the required information:

```
docker save tfg-project:latest | gzip tfg-project.tar.gz
```

Which will export the Docker image from the local installation, streaming its content as an uncompressed tarball, which is piped to GZip in order to provide a compressed version of the image, which can now be imported into another computer using:

```
docker load -i tfg-project.tar.gz
```

The project can then be run using docker run, by mounting a volume with the location of the datasets:

```
docker run -v $(pwd)/datasets:/app/datasets tfg-project <arguments>
```

5.6.4 Usage

The program takes an input file, an output file, an optional training dataset, and the prediction strategy to use as its arguments.

- –**input**: the input file in CSV format. If no training data is provided, the full file is analyzed. If there is training data provided, this parameter will be the live data stream to use. If this parameter is not included, standard input will be considered as input.
- –**output**: output file for the logging of operations. If this parameter is not included, standard output will be considered as output.
- –**train**: training dataset (optional).
- **strategy**: any of the available prediction strategies (OLS, AR...)

Chapter 6

Testing

In this chapter the results of the testing of the ATS will be detailed. The economical and technical outcome of such tests is shown.

6.1 Methodology for testing

In order to test the profitability of the system, a subset of all the available data is chosen. This subset is further divided into two chunks, one being the training dataset, which allows constructing the initial model, and the other being a "live" dataset.

The program will first learn from the training dataset, and then start operating with the live dataset, reacting to the price changes over time.

An amount of 100 is considered as the resources available for trading, and the system will decide to buy/sell the assets at each time, keeping track of its order book.

Every T moments, if the average of the difference of the next N forecasted values with the current price is positive above a F threshold, a K% percentage of the available cash is used to buy the currency. If the value is negative, a K% percentage of the bought assets are sold. If none applies, no operation is done.

At the end of the process, all assets will be sold at their current price, and the resulting total assets minus 100 will be the profit or loss made during the trading session. This process will be repeated with each of the trading algorithms, and compared accordingly.

6.2 Tests performed

A subset of 1 million registers from the Euro - British Pound exchange rates during August 2015 is taken. The first 100 thousand registers are used for training.

While the system manages to show some activity at very low F thresholds, the size of the data and the variability of the stock isn't enough to provide meaningful output to the research. The time constraints found in the project don't allow for a complete research on its performance as it was initially planned.

In order for a full, comparable and verifiable test to be performed, it would be required to perform a parallel processing of the data, since the performance of the system is too low to allow a real time test to be performed.

Chapter 7

Conclusions

This chapter will show the conclusions of the research and future follow-up work that can be done with this subject.

7.1 Conclusions

Automated Trading Systems show as a commonplace approach to financial market trading, which reduces the impact of human decisions and allows for a longer and larger scale operation. However, the lack of precise data for training such systems and being able to operate in live environments makes the design of these systems challenging.

Several approaches to the design of this systems have been described in this research, being the econometric approaches the chosen option for the designed proof-of-concept. Financial markets have been described as well, ending with the choice of foreign exchange markets for their simplicity in operation.

At the end, this research couldn't conclude with a fully functional system as expected, due mostly to time constraints and technical complexity, but it has shown the complexity of this field of knowledge, and the technical challenges that it features.

7.2 Further work

The shown proof-of-concept has severe lacks that makes it unappropiate for using in real environments. A proper implementation would likely require a large development time and more professional resources.

A more complete system, which would connect with real-time systems to gather information and perform operations in a real environment, would be the logical step to progress this research. Such system could also be implemented into a computer cluster and use other, less straightforward models for the predictions.

References

[Akansu, 2017] Akansu, A. N. (2017). The flash crash: a review. *Journal of Capital Markets Studies*, 1(1):89–100.

[Armano et al., 2005] Armano, G., Marchesi, M., and Murru, A. (2005). A hybrid genetic-neural architecture for stock indexes forecasting. *Information Sciences*, 170(1):3–33. https://www.sciencedirect.com/science/article/pii/S002002550300433X?via%3Dihub.

[Attigeri et al., 2016] Attigeri, G. V., Manohara Pai, M. M., Pai, R. M., and Nayak, A. (2016). Stock market prediction: A big data approach. *35th IEEE Region 10 Conference, TENCON 2015*, 2016-Janua. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84962185156&partnerID=40&md5=c47542775423d3234efde26bf948ed9d.

[BARTOS, 2015] BARTOS, J. (2015). Does Bitcoin follow the hypothesis of efficient market? *International Journal of Economic Sciences*, IV(2):10–23. http://www.iises.net/international-journal-of-economic-sciences/publication-detail-189.

[Beattie, 2018] Beattie, A. (2018). The birth of stock exchanges. *Investopedia*, pages 1–6. http://www.investopedia.com/articles/07/stock-exchange-history.asp%5Cninternal-pdf://0.0.0.242/stock-exchange-history.html.

[Chacon and Straub, 2014] Chacon, S. and Straub, B. (2014). Pro Git. *Pro Git*.

[Docker, 2019] Docker (2019). Docker overview. https://docs.docker.com/engine/docker-overview/.

[Easley et al., 2015] Easley, D., López De Prado, M., and O'Hara, M. (2015). *High-frequency Trading*, volume 15. Risk Books,, London :.

[Financial Times,] Financial Times. Equity Definition. http://lexicon.ft.com/Term?term=equity.

[Gervais et al., 2014] Gervais, A., Karame, G. O., Capkun, V., and Capkun, S. (2014). Is Bitcoin a Decentralized Currency? *IEEE Security and Privacy*, 12(3):54–60.

[Gorbunova, 2016] Gorbunova, N. (2016). *European Research Studies*, 19:228–249. https://www.ersj.eu/dmdocuments/16_3_A_p13.pdf.

[Investopedia, a] Investopedia. Earnings Per Share. https://www.investopedia.com/terms/e/eps.asp.

[Investopedia, b] Investopedia. Forward Contract. https://www.investopedia.com/terms/f/forwardcontract.asp.

[Investopedia, c] Investopedia. Spot Trade. https://www.investopedia.com/terms/s/spottrade.asp.

[Investopedia, 2018] Investopedia (2018). What do the bid and ask prices represent on a stock quote? https://www.investopedia.com/ask/answers/042215/what-do-bid-and-ask-prices-represent-stock-quote.asp.

[Investopedia, 2019] Investopedia (2019). Bid-Ask Spread Definition. https://www.investopedia.com/terms/b/bid-askspread.asp.

[Kara et al., 2011] Kara, Y., Acar Boyacioglu, M., and Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, 38(5):5311–5319.

[Kirchgässner et al., 2013] Kirchgässner, G., Wolters, J., and Hassler, U. (2013). *Introduction to Modern Time Series Analysis Second Edition*. Springer Texts in Business and Economics. Springer Berlin Heidelberg, Berlin, Heidelberg. http://link.springer.com/10.1007/978-3-642-33436-8.

[Navale et al., 2016] Navale, P. G. S., Dudhwala, N., Jadhav, K., Gabda, P., and Vihangam, B. K. (2016). Prediction of Stock Market Using Data Mining and Artificial Intelligence. *International Journal of Computer Applications*, 6(6):6539–6544.

http://www.ijcaonline.org/research/volume134/number12/navale-2016-ijca-907635.pdf.

[Pepperstone,] Pepperstone. Foreign Exchange Historical Tick Data. https://pepperstone.com/uk/client-resources/historical-tick-data.

[Shobana and Umamakeswari, 2016] Shobana, T. and Umamakeswari, A. (2016). A Review on Prediction of Stock Market using Various Methods in the Field of Data Mining. *Indian Journal of Science and Technology*, 9(48):1–6. http://www.indjst.org/index.php/indjst/article/view/107985.

[U.S. Securities and Exchange Commission, 2019] U.S. Securities and Exchange Commission (2019). Types of Orders. https://www.investor.gov/introduction-investing/basics/how-market-works/types-orders.

[Virtualenv,] Virtualenv. Virtualenv documentation. https://virtualenv.pypa.io/en/latest/.

[Yermack, 2015] Yermack, D. (2015). Is Bitcoin a Real Currency? An Economic Appraisal. In *Handbook of Digital Currency: Bitcoin, Innovation, Financial Instruments, and Big Data*, pages 31–43.